LeA 35730

Method and computer system for designing experiments

The invention relates to a method and a computer system for designing experiments, and to a corresponding computer program product.

From the prior art, it is known to design experiments by means of statistical experiment designing methods. Such designing methods are used, inter alia, to determine, with a minimum number of experiments, an empirical process model for the relationship between the controlled variables and influencing variables in a process and the resulting product properties and process properties. Such statistical experiment designing can be carried out, for example, using the "STAVEX" (STAtistical experiment designing with EXpert system produced by AICOS Technologie, Switzerland) computer program. A further commercially available computer program for experiment designing is the "Statistica®" program made by StatSoft (Europe) GmbH, Germany.

In the field of statistical experiment designing, various experiment designing types are distinguished in the prior art. In particular, a distinction is made between the classic, fully factorial method and modern methods according to Taguchi or Shainin.

The classic, fully factorial method is the origin of all statistical experiment designing methods. It is based on a comparison of all the quality-conditioned factors with one another by analogy with variance analysis. Numerous variants have been produced over the course of the last few decades and validated in research and development laboratories.

The Shainin DOE (Design of Experiment) is a suitable process for process optimization because it isolates what are referred to as "strong" influencing variables and investigates them for relevance and dependence.

The Taguchi DOE is based on preceding, fractional factorial, orthogonal experiment designs. Because of the drastic savings in terms of experiment runs by preselecting the most important influencing variables, this is a rapid and relatively economic method of designing experiments and processes.

Further known statistical experiment design types of fractional factorial experiment designs, Plackett-Burmann experiment designs, central composite designs, box-Behnken experiment designs, D-optimal designs, mixed designs, balanced block

designs, Latin squares, desperado designs (cf. in this respect also Eberhard Scheffler, Statistische Versuchsplanung and - auswertung; ["statistical experiment design and evaluation"]; Deutscher Verlag für Grundstoffindustrie, Stuttgart, 1997).

Further methods for designing experiments are known from Hans Bendemer, "Optimale Versuchsplanung" [Optimum experiment design], Reihe Deutsche Taschenbücher (DTB, Volume 23, and ISBN 3-87144-278-X) and Wilhem Kleppmann, Taschenbuch Versuchsplanung, "Produkte und Prozesse optimieren" [Optimize products and processes], 2nd expanded edition, ISBN: 3-446-21615-4. These methods are often used in practice for reasons of cost.

The disadvantage with known statistical methods for designing experiments is that the experiment designing and modelling is carried out without taking into account additional knowledge so that, under certain circumstances, no suitable optima are found and the reliability of the results and statements which are generated is questionable. A further significant disadvantage of previously known methods for designing experiments is that when there is a large number of influencing variables to be taken into account, said methods become too extensive. In addition, with certain systems, for example in catalysis or active ingredient research, the target function is often heavily "fractured" and is therefore difficult to capture with statistical methods.

WO 00/15341 discloses a method for developing solid catalysts for heterogeneous catalysed reaction processes, which is based on parallelized testing according to evolutionary methods.

Corresponding methods which operate in an evolutionary way are also known from WO 00/43411, J. chem. Inf. Compute. Sci. 2000, 40, 981-987 "Heterogeneous Catalyst Design Using Stochastic Optimization Algorithms" and from Applied Catalysis A: General 200 (2000) 63-77 "An evolutionary approach in the combinatorial selection and optimization of catalytic materials".

In addition, US 6,009,379 discloses a method for controlling a manufacturing process by means of an efficient experimental design. Here, test points are distributed uniformly on a multidimensional spherical surface in order to be able to weight the individual manufacturing parameters uniformly.

Figure 1 shows a block diagram of a system, known from the prior art, for carrying out screening experiments such as is used in particular in the fields of catalysis and

material and active ingredient research. The system includes a substance library, that is to say what is referred to as a combinatorial library 1 and an experiment set-up 2 for carrying out high throughput screening (HTS) or high speed experimentation (HSE) experiments. Such screening experiments are typically used for identifying active ingredients, catalysis research (homogeneous and heterogeneous), materials research and identification of optimum reaction conditions in chemical, biochemical or biotechnical systems.

A plurality of experiments are usually carried out in parallel in such an experiment set-up 2. The experiment results are output in the form of a file 3. This output data, or some of it, is at the same time the input data for an optimizer 4.

The optimizer 4 is what is referred to as a black-box optimizersr, that is to say an optimizer which is based on a data-driven model or on an evolutionary algorithm. A priori knowledge of the structure and/or interactions is not present in the optimizer 4; instead said optimizer 4 is restricted to the evaluation of the data as such in order to make a selection of experiments from the combinatorial library 1.

The optimizer 4 typically uses the experiment data 3 composed of influencing variables (attributes, factors, structure features, descriptors, physical variables, properties of materials) and data relating to the effect of these variables on what are referred to as targets (target variables), in order to define an optimum search direction within the space of the targets.

Such a black-box optimizer 4 is implemented, for example, by means of:

- genetic algorithms,
- evolutionary algorithms or strategies,
- neutral networks or
- other data-driven model approaches which rely on stochastic or deterministic optimization structures or optimization structures which are a combination of both of these.

A common disadvantage of such systems known from the prior art is that a priori information cannot have an influence, or can only have a restricted influence, in the black-box optimizer 4, and corresponding search strategies often converge slowly, or converge on unsuitable suboptima. Such methods which are known from the prior art are therefore often inefficient in terms of the expenditure of time and the financial outlay. With techniques based on evolutionary algorithms, there is also the risk of the

expenditure and outlay being higher when the optimizer is used to reach the optimum than when a rational or statistical procedure is used.

The invention is therefore based on the object of providing an improved method for designing experiments and a corresponding computer system and computer program product.

The object on which the invention is based is respectively achieved by means of the features of the independent patent claims. Preferred embodiments of the invention are given in the dependent patent claims.

The subject matter of the invention is a method for designing experiments for achieving an optimization goal having the following steps:

A) selection of at least a first experiment from an experimental space by means of a data-driven optimizer in a computer unit,

B) inputting of experimentally determined experiment data of the first experiment in at least one meta layer into a computer unit,

C) use of at least one meta layer for the evaluation of the experiment data,

D) inputting of the experimentally determined experiment data of the first experiment into the data-driven optimizer,

E) influencing of the data-driven optimizer by the result of the evaluation in the meta layer and checking the goal achieved,

F) selection of at least a second experiment from the experimental space by means of the data-driven optimizer,

G) repetition of steps B) to E) for the data of the second experiment,

and

H) stopping the hexation on achieving the goal or repeating steps A) to F) for at least a third or subsequent experiments until the goal has been achieved.

The method is repeated until the optimization goal has been achieved or until it is concluded that it may not be possible to achieve the optimization goal. The method can be terminated automatically or by the user. The optimization goal may be to reach certain evaluation characteristic numbers for the experiments. The characteristic numbers may, for example, be yield selectivities, space-time yields, costs, physical properties, action mechanisms, derived properties, etc. It is also possible to evaluate the experiments using a plurality of characteristic numbers.

The invention permits knowledge for influencing the black-box optimizer to be integrated with the objective of speeding up the convergence and/or ensuring convergence at a suitable optimum as well as increasing the reliability of the results. The knowledge may be known here a priori as prior knowledge and/or may be supplemented continuously by evaluating experiments which have been carried out previously.

Additional knowledge is preferably generated here in the form of "rules", in particular relating to the structure-interaction with data mining and other methods. These rules can be integrated in the designing of the experiment, before, during or after an optimization step or even continuously, the data-driven optimizer being influenced correspondingly. A meta layer is provided for influencing the data-driven optimizer.

The black-box optimizer is tuned by using such a meta layer. In this context, the meta layer is not restricted to one method but rather may contain a combination of various methods. Possible methods are:

- neural networks,
- hybrid model,
- rigorous models,
- data mining methods, for example decision tree methods, general separation methods, subgroup search methods, general partition methods, cluster methods, association rule generators and correlation methods.

The method of operation of the optimizer can be influenced directly here by intervening in the method of operation of the optimizer, or indirectly by filtering the data which form the basis for the optimization.

According to one preferred embodiment of the invention, methods for influencing the optimizer are used which tune the optimizer and/or the optimization process.

Such methods include, for example, subgroup search methods or correlation analyses or attribute statistics in the case of rule generators.

According to one further preferred embodiment of the invention, further meta layers are provided which improve the respectively preceding meta layer or intervene in the preceding meta layer or layers and/or also intervene directly in the black-box optimization process of the first level.

According to a further preferred embodiment of the invention, the intervention positions in the original optimization process and the methods or combinations of methods which are used in the meta layer or layers can be varied in each optimization step. The selection of suitable methods for generating optimum rules can be carried out automatically here.

According to one preferred embodiment of the invention, the optimizer is influenced by a re-evaluation of the experiment data. For example, the experiment data itself can already contain an evaluation by virtue of the fact that appropriate data, for example the yield, is determined directly by experimental means. In this case, the re-evaluation can be carried out by filtering the yield data, for example by virtue of the fact that particularly good yields are given a heavier weighting by means of the data filtering, and particularly bad yields are given a lighter weighting by means of the data filtering. A more rapid convergence of the experiment sequence can be achieved by the means of this type of data filtering.

A corresponding procedure can be adopted if the experiment data does not directly contain an experimentally determined evaluation but rather the evaluation is determined only by means of calculations which follow the experiment. In this case, filtering or weighting is performed not on data which is determined experimentally but rather on evaluations which are determined by calculation.

The method of filtering results here from rules or other relationships which have been found on the basis of an analytical method of the experiment data, for example by means of neutral networks or data mining methods or other methods.

According to a further preferred embodiment, the optimizer is influenced by reducing, enlarging and/or displacing the experimental space.

According to a further preferred embodiment, the filtering is carried out by means of preselection and/or weighting of the experiment data. Particularly "bad" experiment

data, that is to say experiment data which has been recognized as unsuitable by, for example, a rule generator, is preselected and eliminated from the experimental space. In addition, entire columns or rows can also be eliminated from the experiment data matrix if the corresponding parameters have been recognized as irrelevant by the rule generator. As a result, the experimental space is reduced, which considerably reduces the overall expenditure.

The experiment data can be weighted in that experiment data which is recognized as being particularly relevant is duplicated a single time or repeatedly in the experiment data matrix. Alternatively, a weighting coefficient can be introduced.

According to one further preferred embodiment of the invention, the black-box optimizer contains what are referred to as core modules or core operators as well as a model for selecting new test points. The method of operation of the optimizer is then influenced by influencing the core module or modules and/or the module for selecting new test points based on relationships which have been recognized by, for example, a rule generator.

Preferred embodiments of the invention will be explained in more detail below with reference to the drawings, in which:

Figure 1 shows a block diagram representing a system for designing experiments which is known from the prior art,

Figure 2 shows a block diagram of an embodiment of a system according to the invention for designing experiments,

Figure 3 shows a block diagram of an embodiment of the system according to the invention for designing experiments with a re-evaluation of the experiment data,

Figure 4 shows an embodiment of the system according to the invention for designing experiments with preselection and/or weighting of the experiment data,

Figure 5 shows an embodiment of the system according to the invention for designing experiments with influencing of the selection of new test points of the optimizer,

Figure 6 shows an embodiment of the system according to the invention for designing experiments with influencing of the core module or core modules of the optimizer.

The system for designing experiments in Figure 2 is based on a combinatorial library 5 which is formed on the basis of the peripheral conditions given by means of an experimental space. From this combinatorial library 5, an optimizer 6 selects one or more experiments which are then carried out in an experiment set-up 7, for example by means of a high throughput screening or high speed experimentation experiment method. The corresponding experiment data is output in the form of a file 8.

In the system for designing experiments, a meta layer 9 is provided for the optimizer 6. The meta layer 9 is used to influence the optimizer 6 taking into account a priori knowledge or knowledge acquired while the experiment is being carried out. Knowledge, for example in the form of rules or in the form of trained neural networks, can be acquired here continuously by the evaluation of files 8.

The meta layer 9 therefore complements the data-driven optimizer 6 by means of additional knowledge in order to speed up convergence of the experiment series. The meta layer 9 therefore also permits the convergence speed of a black-box optimization method, which is implemented in the optimizer 6, to be improved by integrating prior knowledge and/or rule structures.

This integration can be carried out in various ways, for example by means of:

A information-supported additional selection of the test ensembles, i.e. restriction of the combinatorial library to be tested by means of the rules found with data mining and no intervention into the optimizer

B selective weighting of the optimization steps in the direction of library areas identified as optimum, i.e. intervention into the search method of the optimizer,

C tuning of the selection rules of the black-box optimization methods, i.e. direct intervention into the evaluation method of the optimizer or modification of the evaluation variables before being input into the optimizer

The forms of intervention A, B and C may basically also be carried out in combination, i.e. in an optimization step it is also possible for interventions to be carried out with A and B, B and C, A and C or A and B and C. The intervention positions and intervention combinations as well as the methods used in the meta

layer may change from optimization step to optimization step. The interventions can also be carried out from subsequent meta layers.

When optimizing by means of statistical experiment design, the procedure is similar to the use of a black-box optimizer, that is to say here too intervention is carried out in the optimization process by means of the meta layer in one or more of the forms described above. For example, the integration of prior knowledge is carried out by virtue of the fact that when the influencing variables are selected their field of validity and/or additional restrictions of the field of validity are included in the combination of influencing variables.

Further information on influencing variables may be included for the sequential statistical designing of experiments by using data mining methods or other methods described above, and integrated into the designing of experiments, that is to say the experimental space is changed on the basis of the additional information after the first, second ..., n-th path, respectively.

The change is carried out by

a) adding or removing influencing variables
b) changing the fields of validity of the individual influencing variables or combined influencing variables
c) combination of a) and b).

It is particularly advantageously here that "classic" methods for designing experiments which are known from the prior art can continue to be used for a black-box or a statistical optimizer 6; these methods for designing experiments are improved by means of the present invention by virtue of the fact that taking into account prior knowledge or knowledge acquired during the experiment sequence speeds up the convergence of the experiments or actually permits the convergence of the experiments per se.

In particular, the convergence speed is considerably increased by the tuning according to the invention when optimizing the designing of experiments for catalysts, active ingredients or materials or reaction conditions. A further advantage is that the number of experiments can be reduced while the same results can be expected, making possible the lower degree of expenditure in terms of time and materials and better utilization of the systems.

It is also of particular advantage that integrating the prior knowledge prevents loss of research investment when HSE or HTS technologies are used or in a combinatorial procedure.

Figure 3 shows an embodiment of the system for designing experiments in which the experiments are re-evaluated.

One or more experiments which have been previously selected from the combinatorial library 5 (cf. Figure 2) are carried out in the experiment set-up 7. The corresponding experiment data is output in the form of the file 8. The experiment data may itself already contain an evaluation here if appropriate data can be acquired directly by experimental means. An example of this is the experimental determination of the yield. The yield is at the same time an evaluation of the experiments carried out.

In other cases it may be necessary for an evaluation of the experiment data to be additionally performed in an evaluation module 10. For example, the evaluation module 10 contains a calculation rule for the calculation of an evaluation based on one or more of the experiment data.

The file 8 and, if appropriate, the result of the evaluation module 10 are input into the meta layer 9. The meta layer 9 contains a module 11 for implementing a data mining (DM) algorithm, a neural network or a hybrid method or some other suitable data analysis method.

Rules are generated by applying such a method, that is to say additional information and observations relating to the understanding of the chemical system considered in the experiments. The module 11 therefore has the function of a rule generator. Corresponding rules and secondary conditions are formulated in the module 12 of the meta layer 9.

A re-evaluation of the experiment or experiments is then carried out, if appropriate, in the module 13 on the basis of these rules and secondary conditions. This can be carried out in such a way that a re-evaluation of an experiment is carried out only if a predefined threshold value is exceeded. Alternatively, the user can also intervene in order to activate or deactivate the re-evaluation. The re-evaluation may consist in experiments which are recognized as being "poor" are given a worse evaluation and experiments which are recognized as being "good" are given an improved evaluation.

On the basis of the file 8, which, if appropriate, contains re-evaluated experiment data, the black-box optimizer 6 then creates a further experiment design 18. The corresponding experiments are then in turn carried out in the experiment set-up, and so on.

Figure 4 shows an alternative embodiment in which the filtering is not carried out by means of a re-evaluation of the experiment data, but rather by a preselection and/or weighting. The system for designing experiments in Figure 4 has basically the same design here as that in Figure 3, a module 15 for the preselection and/or weighting being used instead of the module 13.

The experiment data is therefore not re-evaluated or evaluated differently, but instead the module 15 can be used, for example, to eliminate experiments or give them greater or lesser weighting on the basis of the rules determined. As a result, a preselection takes place without the actual evaluation of the experiments being changed.

Figure 5 shows a further embodiment of the system according to the invention for designing experiments. The embodiment in Figure 5 differs from the embodiment in Figure 3 and Figure 4 in that there is direct intervention into the optimizer 6.

In this embodiment, the optimizer contains one or more core modules 16, that is to say what are referred to as core operators. In addition, the optimizer 6 contains a module 17 for selecting new test points. The method of operation of the module 17 is influenced by the rules and secondary conditions formulated by the module 12, that is to say, for example, new test points selected by the module 17 are rejected so that there is a feedback from the module 17 to the core module 16 in order to select further, corresponding test points as replacements for the rejected test points.

After the actual optimization has occurred in the core module 16 or core modules 16, new experiments or test points for optimizing the target variables of the system under consideration are therefore proposed by means of the module 17. This system may be, for example, a chemical, biotechnological, biological or enzymatic system.

Experiments which contradict the generated rules are eliminated on the basis of the rules produced by means of the rule generator, that is to say the meta layer 9, and if appropriate said experiments are supplemented with new experiments relating to the optimizer, that is to say the core module 16 or core modules 16.

The elimination may be carried out here in a strict way, that is to say completely, or in a soft way, that is to say with a certain degree of weighting. The newly designed experiments must then also in turn pass through the module 17. This ensures that information which is not, or cannot, be taken into account by the core module 16 or core modules 16, is subsequently integrated into the designing of experiments.

Alternatively, a separate module 18 can follow the optimizer 6 in order to perform the post-selection of the new test points selected by the module 17. This corresponds to a test in the module 18 to determine whether the new test points, which have been proposed by the module 17, conform to the rules. If test points are eliminated in this test, feedback is in turn necessary in order to design corresponding alternative new test points.

The embodiment of the system for designing experiments in Figure 6 corresponds to the system for designing experiments in Figure 5 with the difference that the method of operation of the module 17 is not influenced, nor does a post-selection take place in the module 18, but rather the method of operation of the core module 16 or core modules 16 of the optimizer 6 is influenced directly.

Examples of core operators of neural networks are the type and number of influencing variables and the weighting of individual data points.

Examples of core operators of evolutionary algorithms taking the example of the genetic algorithm are the selection operator (selection of a new series of experiments), the mutation operator and the cross-over operator.

The rules and information which are generated by the rule generator are taken into account in the execution in the algorithm of the actual optimizer.

For optimizers which are coupled to neural networks, this means that the experimental space is restricted by the rules, or the data records are weighted in a particular way.

With evolutionary algorithm optimizers, the additional information is taken into account in one or more core operators. This means that, for example, specific cross-overs, selections or mutations are prohibited or carried out with preference.

For both types of optimizer, the result of this is that when there is complete automation of the workflow, there is intervention into the corresponding program

parts of the optimizer via interfaces, or the information is included in the optimizer by means of manual or program-controlled changes of optimization parameters.

The embodiments in Figures 3 to 6 can be combined with one another, that is to say a plurality of rule generators, that is to say meta layers 9, can be integrated into the optimization sequence independently of one another. These rules can be generated using various methods which are used independently of one another, and combined in the module 12.

The rules which have been formulated by the rule generator or rule generators of the meta layers 9 are taken into account either automatically via defined interfaces and with compliance with predefined threshold values, or by means of manual formulation of rules for the area of the optimizer into which the rule generator intervenes.

List of reference numerals

Combinatorial library	1
Experiment set-up	2
File	3
Optimizer	4
Combinatorial library	5
Optimizer	6
Experiment set-up	7
File	8
Meta layer	9
Evaluation module	10
Module	11
Module	12
Module	13
Experiment design	14
Module	15
Core module	16
Module	17
Module	18